图书情报工作
第65卷 第16期 2021年8月

融合语义联想和BERT的图情领域SAO短文本分类研究*

张玉洁[1] 白如江[1] 刘明月[1] 于纯良[2]
[1] 山东理工大学信息管理研究院 淄博 255049 [2]烟台大学图书馆 烟台 264005

摘 要：[目的/意义]针对SAO结构短文本分类时面临的语义特征短缺和领域知识不足问题，提出一种融合语义联想和BERT的SAO分类方法，以期提高短文本分类效果。[方法/过程]以图情领域SAO短文本为数据源，首先设计了一种包含"扩展－重构－降噪"三环节的语义联想方案，即通过语义扩展和SAO重构延展SAO语义信息，通过语义降噪解决扩展后的噪声干扰问题；然后利用BERT模型对语义联想后的SAO短文本进行训练；最后在分类部分实现自动分类。[结果/结论]在分别对比了不同联想值、学习率和分类器后，实验结果表明当联想值为10、学习率为4e－5时SAO短文本分类效果达到最优，平均F1值为0.852 2，与SVM、LSTM和单纯的BERT相比，F1值分别提高了0.103 1、0.153 8和0.140 5。

关键词：SAO 短文本分类 语义联想 BERT
分类号：TP391 G250
DOI：10.13266/j.issn.0252－3116.2021.16.013

1 引言

SAO（Subject-Action-Object）是以特定方法从论文、专利中抽取的能够表达关键概念、关键方法的特定结构的短文本[1]，是文本细粒度表达方式之一。SAO由主体（Subject）、行为（Action）和客体（Object）三部分组成[2]，因其语法结构完整，相较于单一的关键词更能表达丰富的含义，目前在潜在创新点挖掘[3－5]、专利特征分析[6－9]、新兴技术预测[10－12]等方面具有广泛应用。SAO短文本分类研究有利于系统地梳理科学技术发展脉络，高效地实现文本挖掘工作，但是，当前在提高SAO短文本自动分类效果上，仍面临一些阻碍，如何将数量庞大而类目零散的SAO文本进行有效地分类和组织已经成为目前亟待解决的问题。

与普通的长文档和短文本相比，虽然SAO语法结构完整，但表征能力有限、领域专指性弱，可供提取的特征只有Subject和Object以及二者的对应关系Action，故难以得到有效的特征词；同时受限于表达结构，SAO在面向特定领域分析时常常面临领域知识不足问题。因此，本文提出融合语义联想和BERT的SAO短文本分类方法，并以图书情报领域SAO短文本（以下简称为图情SAO）为数据源进行实证，旨在丰富SAO短文本的表征能力，以此解决SAO分类时语义特征短缺和领域知识不足的问题，提高分类性能。

2 相关研究

SAO本质是包含主谓宾的三元组（Triple），由两个节点（Node）及其关系（Edge）组成，是构建知识图谱的基本元素。与SAO相类似的结构还有SPO（Subject-Predicate-Object）和SVO（Subject-Verb-Object）[6]，这些概念均是通过实体识别、句法分析来辨别句子中的句法结构和依存关系，以此提取主谓宾元素。SAO与SPO、SVO的不同之处在于应用场景和使用领域的不同，在谓词选择上各有侧重，其中，近年来SAO在知识挖掘与发现、潜在创新点挖掘、专利特征分析等诸多方面均有广泛应用。

胡正银等从微观层面的SAO构建了语义TRIZ的方法、流程与关键技术，并以大口径光学元件专利为例构建领域个性化语义TRIZ，结果显示提出的方法能有效地实现半自动构建领域个性化语义TRIZ[13]；另外，

* 本文系山东省高等学校青创科技支持计划"科技大数据驱动的智慧决策支持创新团队－面向新旧动能转换的新兴科学研究前沿识别研究"（项目编号：2019RWG033）和山东省社科规划处项目"数字环境下科学论文的内容标注模型研究"（项目编号：20CSDJ65）研究成果之一。
作者简介：张玉洁（ORCID：0000－0002－6819－031X），硕士研究生；白如江（ORCID：0000－0003－3822－8484），研究馆员，硕士生导师，通讯作者，E-mail：brj@sdut.edu.cn；刘明月（ORCID：0000－0002－4335－9369），硕士研究生；于纯良（ORCID：0000－0002－3013－8022），副研究馆员。
收稿日期：2021－01－27 修回日期：2021－05－12 本文起止页码：118－129 本文责任编辑：杜杏叶

该作者后续又基于 SAO 三元组与简单知识对象,融合文本挖掘技术构建细粒度、多维度的领域技术索引,实现了领域知识棱镜、面向 TRIZ 的语义检索与专利可视化分析功能[7]。汪雪峰等[3]通过分析 SAO 三元组提出了基于解决方案相似性来确定研发合作伙伴的方案,并以太阳能电池行业为例进行实证,案例表明该方案能够帮助公司理解研究目标之间的关系;后又在 2019 年提出了基于 SAO 结构的创新解决方案遴选研究[5],该研究以目标研究领域具体研究问题为出发点,在全领域寻找潜在解决方案,并从技术可行性以及预期效果两方面对这些潜在解决方案进行评价,实证表明该方法具备有效性。

SAO 分类可以视为自然语言处理(Natural Language Processing, NLP)领域一种特殊的文本分类类型,即短文本分类,其本质是将文本内容转换成机器可识别的向量化表示,通过机器自动学习文本特征来识别不同的类别。但是,与长文档分类不同,短文本字数稀少、特征稀疏,常规方法难以捕获有效特征,因此大多数学者围绕机器学习、深度学习、混合模型、预训练模型及数据扩充等方面展开研究。

传统机器学习的短文本分类方法诸如支持向量机[14]、贝叶斯[15]、隐马尔可夫模型[16]、随机森林[17]等是通过对样本数据构造特征工程,再输入特定的分类器实现训练和预测[18－19]。但该类方法前期需要构造大量特征工程,泛化能力弱,无法充分利用大规模数据学习特征[20]。深度学习的短文本分类方法克服了传统机器学习的缺陷,它着重于模型构建和参数调整,通过深层次的非线性变换在大量训练数据上拟合特征值,诸如循环神经网络[21]及其变种模型[22－24]、卷积神经网络[21]、注意力机制[25]等深度学习方法均在短文本分类上有良好表现。邓三鸿等[26]融合长短期记忆网络模型和字嵌入方法对中文图书标签进行分类,通过题名、主题词等短文本特征训练模型,在 3 所高校的 5 个类别书目数据的分类实验上取得良好效果;赵亚娟[27]、Franck[28]等分别利用循环神经网络、卷积神经网络及混合方法对专利、对话行为等领域的短文本数据进行分类;章成志[29]、陶志勇[30]、余本功[31]等或改进或融合的层次注意力网络,为短文本的特征表示相关工作中提供了许多研究思路。但这些方法缺乏对文本深层次含义的发散,同时需要大量有标签数据进行训练,对数据的数量和质量都有相当高的要求,简单少量的数据难以适应复杂的网络模型。

2018 年谷歌提出 BERT[32]预训练模型,采用多个双向 Transformer[25]结构的编码器,设计大量多头注意力机制(Multi-head Attention),依靠大规模训练数据学习通用知识,辅以少量领域数据进行微调,在包括文本分类的多个下游任务中取得 SOTA(State-Of-The-Art)结果。X. Qiu 等[33]详细对比了 BERT 在文本分类上的各种方法,在微调策略、进一步预训练和多任务训练等多种不同方式提出许多思路;J. S. Lee[34]、X. Lu[35]等使用 BERT 在专利数据分类上进行微调,均实现了较为理想的效果。

在 BERT 融合语义信息和领域知识的研究上,一些学者通过改进 BERT 输入模式,来提升文本语义信息[36]。W. Liu 等[37]提出 K-BERT,将训练数据映射到领域知识三元组中以增加输入数据的领域知识,同时添加一层可视化层用以解决知识噪音问题,在多个数据集上取得了不错表现;S. Yu 等[38]提出为文本构造辅助句和领域知识,把分类任务转换为二进制句子对,探讨了学习策略、学习率、序列长度和隐藏状态向量对分类结果的影响。

上述学者的相关研究为本文提供了重要思路:依托预训练模型,通过对数据集进行具体领域的语义联想,可以缓解短文本语义特征短缺和领域知识不足的问题。但针对特殊的 SAO 结构短文本,尤其是特殊领域如图书情报领域的 SAO 分类,尚未有相关研究进行论证或提出较为理想的解决方案,文章在上述研究基础上,将对 SAO 短文本分类做进一步探究。

3 融合语义联想和 BERT 的 SAO 短文本分类设计

本文研究框架如图 1 所示,主要包括语义联想、融合语义联想的 BERT 和分类三大部分。语义联想旨在提高 SAO 语义表征能力和解决语义噪声干扰问题,BERT 用于微调语义联想后的 SAO 数据,最后在分类部分选择适当的分类器实现短文本自动分类。另外文章还将对比不同模型、联想值、学习率、分类器对短文本分类结果的影响,以期探寻适合于图情 SAO 短文本分类的最佳方案。

3.1 语义联想

本文提出的语义联想方案由语义扩展、SAO 重构和语义降噪三部分组成,其目的是为 SAO 扩展更多上下文信息,在特征编码时捕获更多领域知识,同时防止联想过度导致语义表达与原 SAO 相偏离,因此,该方案包含“扩展－重构－降噪”三个环节。设定输入为 SAO,已经训练好的 Word2Vec[40]图情领域模型记作 M。

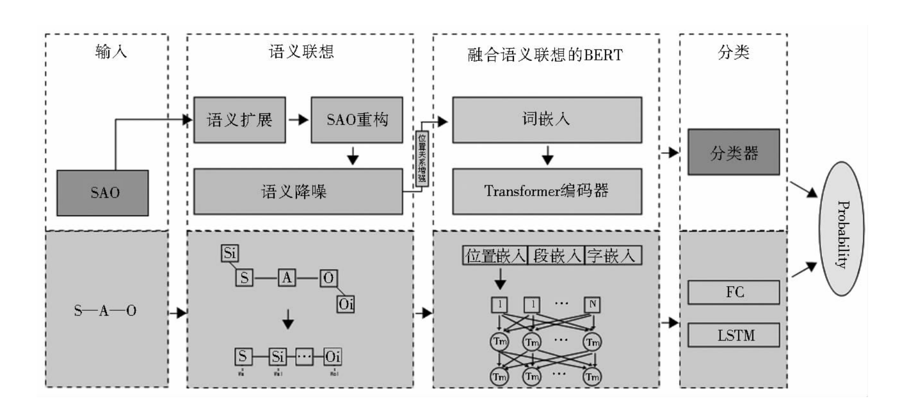


图 1 研究框架

3.1.1 语义扩展

图书情报领域 SAO 与普通 SAO 相比除了语义特征短缺外，还存在领域知识不足的问题。具体来说，在短文本自动分类时常常面临的重要挑战是因图书情报领域的公开标注样本稀少、人工标注成本高昂、知识水平差异而导致的标注质量参差不齐、领域内分类特征的非显著性以及领域编码的稀疏性等问题，因此语义扩展的目的是为 SAO 扩展更多图情领域上下文信息，在特征编码时捕获更多图书情报领域的知识。为此，笔者训练了图书情报领域 Word2Vec 模型，该模型为 Subject 和 Object 分别适配在向量空间中最相近的同义表达，既能丰富 SAO 的表达能力，又能补充更多的领域知识。

在语义扩展时，通过计算余弦距离[39]返回最相似的备选同义表达，余弦距离能够反映两个词在空间位置上的相似性，计算公式如公式(1)所示，其中 X 为目标词汇，Y 为模型空间所有词汇，计算得出的备选词汇与原 SAO 进行映射，挂载到各自对应查询词下，生成扩展后的 SAO 短文本，以树状结构存储为 SAO 树，记作 T，其计算公式如公式(2)所示，其中 n 为联想值。

$$\cos(\theta)=\frac{\sum_{i=1}^{n}(X_i\times Y_i)}{\sqrt{\sum_{i=1}^{n}(X_i)^2}\times\sqrt{\sum_{i=1}^{n}(Y_i)^2}} \quad 公式(1)$$

$$T=\{(S\cdots S_n)A(O\cdots O_n)\} \quad 公式(2)$$

图情 SAO 短文本核心在于 Subject 与 Object，Action 作为谓词只反映 SAO 的主客体关系，因此不对 Action 进行扩展，只对 Subject 和 Object 扩展。如“大学图书馆，构建，学习共享空间”进行 n = 2 的扩展后，其结构见图 2(a)。

3.1.2 SAO 重构

BERT 模型接受序列结构的输入，因此需要将 SAO 树构造成线性序列结构的文本，记作 L。此时有两种重构方案，以“大学图书馆，构建，学习共享空间”为例，方案 1 为{大学图书馆高校图书馆研究型大学图书馆构建学习共享空间信息共享空间实体空间}，见图 2(b)，记作 L1，表达式如公式(3)所示：

$$L_1=\{S_0\cdots S_n\ AO_0\cdots O_n\} \quad 公式(3)$$

方案 2 为{大学图书馆构建学习共享空间，大学图书馆构建信息共享空间，大学图书馆构建实体空间，高校图书馆构建信息共享空间，高校图书馆构建学习共享空间，高校图书馆构建实体空间，研究型大学图书馆构建实体空间，研究型大学图书馆构建学习共享空间，研究型大学图书馆构建信息共享空间}，见图 2(c)，记作 L2，表达式如公式(4)所示：

$$L_2=\left\{[S_0\cdots S_n]\times A\times\begin{bmatrix}O_0\cdots\\ \cdots\\ O_n\end{bmatrix}\right\}=\{S_0\ AO_0,\cdots,S_n\ AO_n\} \quad 公式(4)$$

两种方案重构的 SAO 均存在与原 SAO 表达含义相偏离的问题，即语义噪声。L_1 的语义噪声在于丢失语法结构关系，原本的主谓宾关系经过重构后无法表达完整的语义，编码的词间关系错位，导致获取到错误前后文信息；L_2 的语义噪声在于扩展的 SAO 搭配过载，造成过度联想，导致扩展结果与原本的表意相悖。针对该问题，本文提出语义降噪解决方案。

3.1.3 语义降噪

由于训练语料的差异化分布和词向量表示过程的黑盒属性，Word2Vec 为图情 SAO 的语义扩展难免存在不相关甚至相悖的特征词，因此扩展和重构后的图情 SAO 需要进一步“清洗”。语义降噪的目的是降低语义联想后的 SAO 对原 SAO 的噪声干扰，同时最大程度保留联想信息。基于此，笔者借鉴注意力机制的核心思想，为每个扩展词进行“打分”，对语义联想后的 SAO 有选择性地挂载或遗忘，突出重点、舍弃冗余。

具体思路是：为每个扩展词分别赋予权重，权值以

目标词汇与检索所得词汇的 Word2Vec 相似度表示，不同权重代表不同扩展词的重要程度，原始 SAO 各自权重均为 1，随后对 L_2 中每个 SAO 进行加权求和，排序后自顶向下取 n+1 个 SAO 作为语义联想后的 SAO，即 L，表达式见公式(5)，其中 w 表示不同扩展词所赋予的不同权值。

$$L=\left\{\sqrt[2]{\sum_{i=0}^{n} wS_0\ wAwO_0,\cdots,\sum_{i=0}^{n} wS_n\ wAwO_n}+1\right\}$$

公式(5)

如图 2(d)所示，“大学图书馆，构建，学习共享空间”语义降噪后的 SAO 表达为{大学图书馆构建学习共享空间(3)，大学图书馆构建信息共享空间(2.781)，高校图书馆构建学习共享空间(2.735)，大学图书馆构建实体空间(2.717)}。降噪后的 SAO 在语义联想的基础上降低了噪声干扰，最大程度地保证了语义完整性和发散性，保留了 SAO 结构的位置关系，在 BERT 词嵌入时保证了位置嵌入的可解释性。

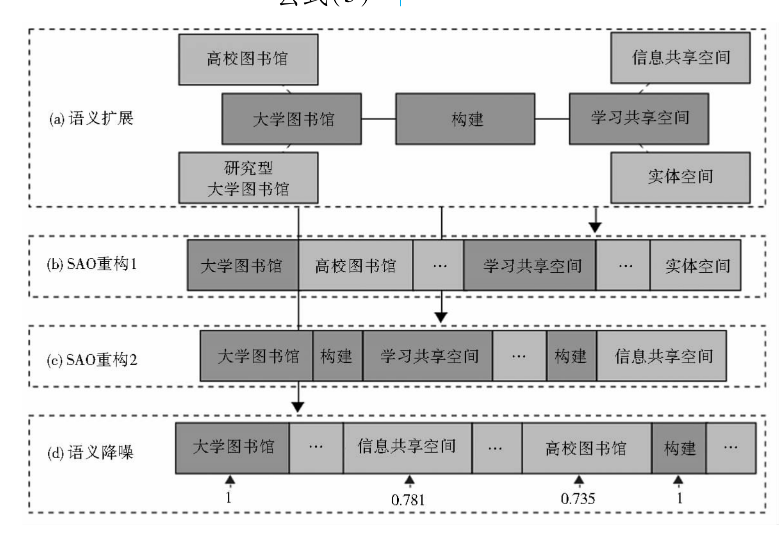


图 2　语义联想示例

3.2　融合语义联想的 BERT

SAO 经过语义联想后输入 BERT 实现微调和训练，本文选择 BERT 是因为 BERT 作为预训练模型，本身就已经在大量数据训练的基础上携带通用领域的先验信息。采用 BERT 进行图情 SAO 分类，只需要微调语义联想后的 SAO 数据，从而缓解重新开始训练复杂模型参数所导致的过拟合或欠拟合问题，以期提高图情 SAO 词向量的表征能力。

BERT 预训练参数结合语义联想后的 SAO 获取的新的训练参数，依次经过词嵌入和多层双向的 Transformer。词嵌入用于将输入文本转换成向量表示，Transformer 通过编码器捕获文本权值信息。

词嵌入主要分为 3 个过程：字嵌入(Token Embedding)、段嵌入(Segment Embedding)和位置嵌入(Position Embedding)，见图 3。字嵌入通过 BERT 字符查询表将 SAO 转换为字符级的一维向量表示，在 MASK 时随机遮罩一部分字符，获取从左向右和从右向左的双向信息，[CLS]用于标记一条 SAO 的开始，[SEP]标记结束；段嵌入标记不同的符号用以获取文本的全局语义信息和识别不同的 SAO，并与字符级的向量相融合；位置嵌入标记前后文信息；最后字嵌入、段嵌入和位置嵌入相加输出最终的词嵌入表示。

BERT 的位置嵌入是区别于其他模型的重要之处，位置信息使每个字对其他字的影响不完全相同，使 BERT 可以根据上下文动态地捕获字词前后的关联性。在位置嵌入时，字符的权重系数并不是由某个固定参数决定，而是由前一个字符计算权重后与该字符权重进行融合，这样在生成下一个字符时，原本固定的参数 w 被替换为根据上一个字而动态变化的参数 wi，SAO 的每一个字符都注入了上一个字符信息，输入越长，权重系数越重要。如图 3(a)所示，原始 SAO 能获得的前后文信息不足，在计算前后文信息时扩展权重系数只能计算到 15，语义联想后的 SAO 的词嵌入如图 3(b)所示，经过联想后，比原始 SAO 权重系数更高，更加注重上下关系的嵌入，模型能捕获到的细节信息更加丰富。

词嵌入后连接 Transformer 编码器。Transformer 作为 BERT 的特征抽取器，采用多层双向的结构计算隐藏状态向量，其中包含 12 个 Transformer 编码器和多头注意力(Multi-Head Attention)，层层累积形成 BERT。

Transformer 编码器经过多头注意力把位置信息加入到编码中,并考虑前一个字对当前字的权重影响,将输入维度和输出维度对照起来,经过归一化后输出隐藏状态向量。



图 3　融合语义联想前后的 BERT 词嵌入比较

3.3　分类

经过 BERT 训练后,SAO 以向量形式表示,在分类部分连接相应的分类器实现自动分类。短文本多分类时通常接入全连接层或其他预置网络模型,最后使用 Softmax 作为激活函数实现分类。

Softmax 实现最终分类预测结果的过程如下:给定包含 i 条 SAO 的集合:

$$D = \{(L^1, label^1), \cdots, (L^i, label^i)\} \quad \text{公式(6)}$$

其中 L^i 代表第 i 条语义联想后的 SAO, $i \in R$, $label^i$ 表示第 i 条 SAO 对应的类别, $label^i \in \{1, 2, \cdots, c\}$, c 为分类数量,对于任意一条 SAO,以期通过 Softmax 函数计算出条件概率分布,即 SAO 属于每个类别的概率,返回 c 维矩阵,其中概率值最大值即为该条 SAO 所属类别。

BERT 层经过训练后输出模型参数 θ 和隐藏状态向量 H,因此 Softmax 目标为计算概率分布 $P(label_i | H[CLS], \theta)$,其公式如(7)所示,经过上述策略后输入 SAO 的分类概率值。

$$P(label_i \mid H_{[CLS]}, L) = \frac{e^{label_i \mid H[CLS], L)}}{\sum_{j=1}^{c} e^{label_j \mid H[CLS], L)}} \quad \text{公式(7)}$$

针对本文语义联想后的图情 SAO 这样多条语句表达相似含义的句式,使用其他文本分类器对分类结果可能会有不同影响,产生不同分类结果,因此本文将对比不同分类器对短文本分类指标的影响。

3.4　评估指标

为评估 SAO 短文本分类效果,本文采用 Precision, Recall 和 F1 值作为评价指标,如公式(8)至(10)所示,P 值常用于评估预测结果中正确占比情况,P 值越高则预测正确率越高,模型效果越好;R 值越高则分类越准确,模型效果越好;通常情况下查全率与查准率无法同时达到高标准,而单纯以 P 值或 R 值作为衡量指标缺乏全面性,因此使用 F1 值取加权调和平均。

$$P = \frac{\text{预测正确结果}}{\text{预测出的所有结果}} \quad \text{公式(8)}$$

$$R = \frac{\text{预测正确结果}}{\text{样本中的所有结果}} \quad \text{公式(9)}$$

$$F1 = \frac{2 * P * R}{P + R} \quad \text{公式(10)}$$

4　实证研究

基于上述设计思路,本部分开展实证研究。为对比本文提出的融合语义联想和 BERT 的 SAO 短文本分类方法和传统机器学习、深度学习之间的差异,实验选择支持向量机(SVM)与长短期记忆网络(LSTM)作为对比基准模型;为比较不同数量的扩展词对分类效果的影响,将进行不同联想值下的实验;为对比不同学习率、分类器因素对结果的影响,将选择联想值分类效果最优的一组分别进行不同学习率与分类器的对比实验。

4.1　实验环境

硬件配置:Intel E5 – 2609v4 + NVIDIA TESLA P4 * 1

软件配置:Win10 + Python3.6 + Tensorflow1.5 + Keras2.1 + PaddlePaddle1.7

4.2　语料来源及数据集

4.2.1　语料来源

本文数据来源于图书情报领域学科《中国图书馆学报》《情报学报》《大学图书馆学报》《图书情报知识》《图书与情报》《情报资料工作》《图书情报工作》《情报理论与实践》《情报杂志》《情报科学》《图书馆论坛》《国家图书馆学刊》《数据分析与知识发现》原

《现代图书情报技术》《图书馆学研究》《图书馆》《图书馆建设》《图书馆杂志》《现代情报》在内的18种CSSCI期刊，每种期刊选取被引频次前500的论文题录信息，包含2000-2018年间的共计9 000条数据，每条论文数据包含题名、作者、关键词等属性。

4.2.2 Word2Vec数据集及模型

Word2Vec数据集用于训练Word2Vec模型，旨在后续对图情SAO实现语义联想。为了提高语义联想词的质量和新颖程度，训练数据集在上述9 000条数据基础上，又增加了《中国图书馆学报》《图书情报工作》《情报学报》《数据分析与知识发现》四种期刊在2000-2020年刊发的11 931条论文题录信息。由于所得数据属性与原9 000条数据属性有所差异，且构建模型对词汇重复性并无要求，因此不进行数据去重，最终得到构建Word2Vec模型的数据共计20 931条。

Word2Vec训练之前需要分词、去停用词、大小写转换、删除无用符号等预处理操作，为了尽可能保证模型质量，笔者从《中国大百科全书 图书馆学·情报学·档案学》[41]和《新编图书馆学情报学辞典》[42]中抽取了60 503个词条作为jieba分词的外部词典，定义了4 652个常见字、词、符号作为停用词表，选择Gensim库实现训练过程，各项参数分别为：维度100，修剪词典数量3，训练算法Skip-gram，跳词窗口5，经过10次迭代后完成训练。如查询与"信息"相近的词语，可视化词向量结果如图4所示：

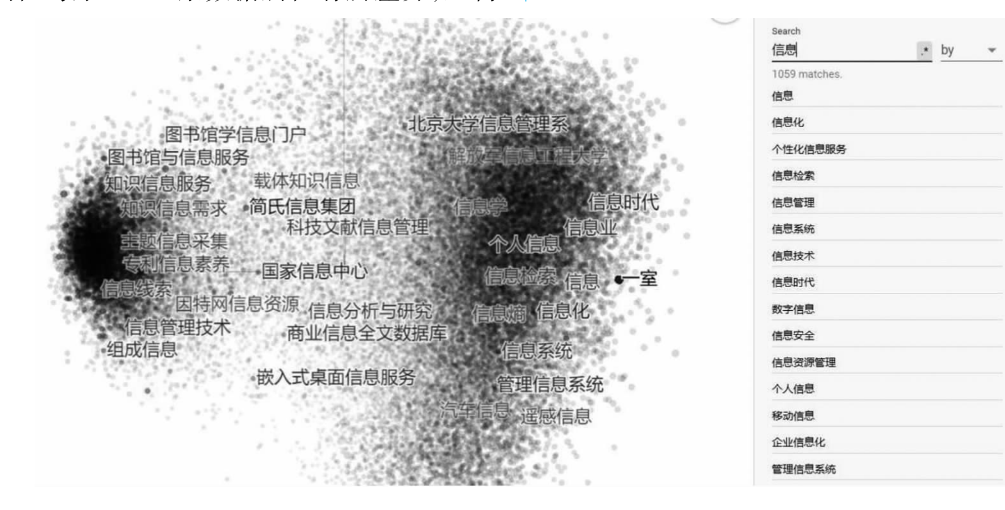


图4 Word2Vec训练词向量可视化

4.2.3 SAO短文本分类数据集

图书情报SAO短文本分类数据集是本文进行自动分类的目标数据，抽取论文数据的字段包括题目和摘要，抽取方法是基于哈工大LTP依存句法分析和语义角色标注的事件开源项目来抽取SAO三元组，不同的是本研究在分词时使用了自定义图情词典和停用词表，在调用程序中的TripleExtractor类方法后，重写程序功能应用到本文的数据上。另外，在抽取后，我们定义了SAO筛选与清洗规则[43]，对SAO的质量进行过滤，从而保证SAO的可用性。清洗后经过人工标注的图书情报领域SAO短文本数据，共计11 021条，每条数据包含subject、action、object和label四项属性，SAO短文本的字符长度分布和词频分布情况见图5。

本文分类标签参考全国技术名词审定委员会公布的《图书馆·情报与文献学名词2019》[44]的八大分类方法，考虑到训练语料分布状况和词向量表示的限制，经查询相关文献、标准、专利和多次专家讨论后定为6大类别，分为信息资源建设、信息组织、图书情报工作管理、信息服务与用户研究、情报分析与研究、其他。数据经过人工标注后，由本领域专家进行意见反馈，经由4位图书情报领域专家和学者多次讨论和修改后最终确认。各类别名称与Id、数量对应关系如表1所示，上述数据按照8:2的比重进行随机抽取，并设置随机种子，确保随机抽样的可控性。

表1 类别-标签-样本量对照

Label	信息资源建设	信息组织	图书情报工作管理	其他	信息服务与用户研究	情报分析与研究
Id	0	1	2	3	4	5
Quantity	1 708	2 326	1 932	1 562	1 968	1 525

4.3 基于基准模型的SAO短文本分类实验

4.3.1 基于SVM的SAO短文本分类实验

支持向量机(SVM)是应用最为广泛的机器学习算法之一，其原理是在N维空间中找到一个超平面对数据点进行划分，使两类别距离该平面的距离最大化，相

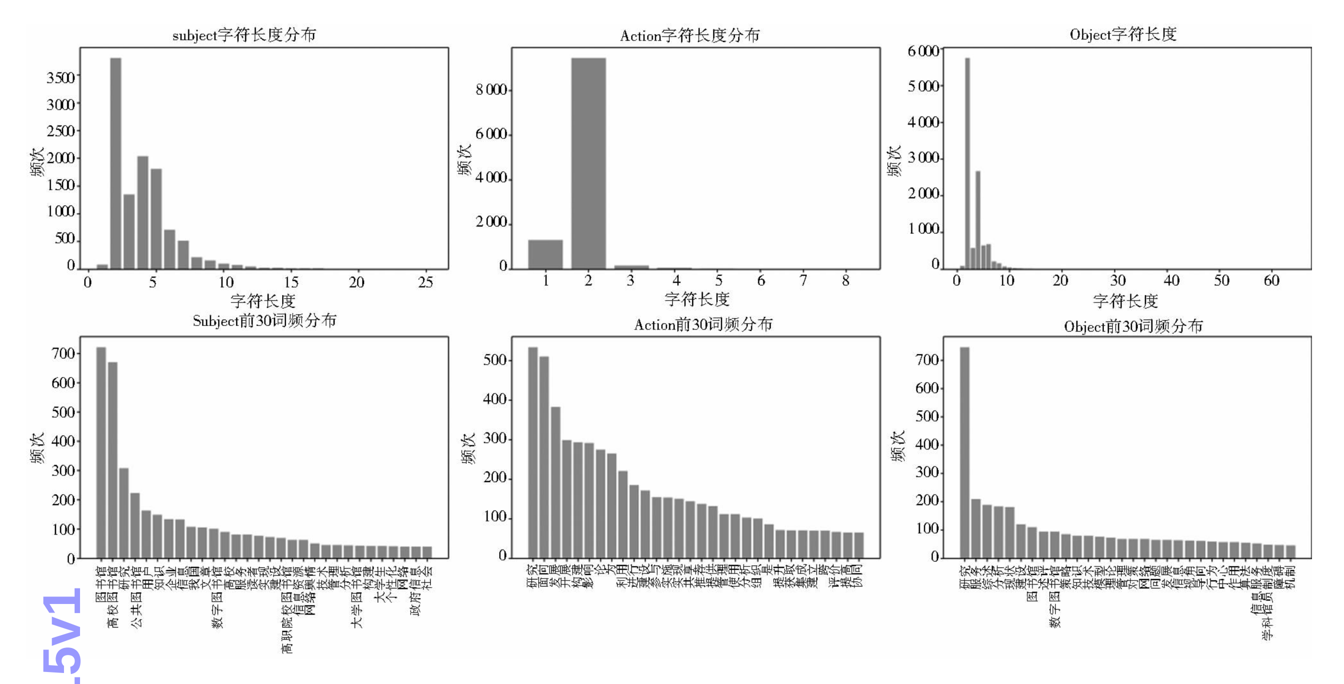


图 5 图情 SAO 语料分布情况

关研究[45-47]中使用 SVM 模型均取得不错的效果，因此本文选择 SVM 作为对比基准模型之一。SVM 经过分词、去停用词等预处理后构造数据特征，依次经过特征选择和特征权重计算，构造 SVM 分类器，经过多次迭代和优化后，准确率达到 0.75，平均 F1 值 0.749 1。各类别参数如表 2 所示：

表 2 SVM 模型各项分类指标

Id	P	R	F1
0	0.752 4	0.701 8	0.726 2
1	0.789 2	0.789 2	0.789 2
2	0.757 0	0.772 6	0.764 7
3	0.737 5	0.711 5	0.724 3
4	0.716 6	0.802 0	0.756 9
5	0.760 6	0.708 2	0.733 4
Average	0.752 2	0.747 6	0.749 1

4.3.2 基于 LSTM 的 SAO 短文本分类实验

长短期记忆网络（LSTM）是循环神经网络（RNN）的变形之一，借助门（Gata）机制降低句子的长期依赖，有效化解了梯度消失及梯度爆炸问题，广泛应用于文本分类问题[26,48]，因此选作本文基准模型之一。LSTM 输入编码映射到词典为每个词分配一个编号后向量化，每条 SAO 转换成一个整数序列的向量，激活函数设置为 Softmax，损失函数设置为分类交叉熵。经过多次训练迭代之后，当 Epochs 为 10，Batch_size 为 32 时效果最优，准确率 0.7，平均 F1 值 0.698 4，各项指标如表 3 所示：

表 3 LSTM 模型各项分类指标

Id	P	R	F1
0	0.645 7	0.602 5	0.623 3
1	0.684 5	0.682 4	0.683 4
2	0.701 5	0.674 5	0.687 7
3	0.683 3	0.737 4	0.709 3
4	0.716 7	0.722 2	0.719 5
5	0.776 6	0.754 2	0.765 3
Average	0.701 4	0.695 5	0.698 4

4.4 融合语义联想和 BERT 的 SAO 短文本分类实验

4.4.1 不同联想值下的 SAO 短文本分类实验

不同扩展词能够生成不同长度和语义的 SAO，为对比联想值对分类效果的影响，本部分进行不同联想值下的实验。综合考虑硬件条件与 Word2Vec 规模大小后，联想值 n 分别设置为 0、5、10、15，模型选择中文版 BERT，四项实验采取统一配置参数：Epochs 为 10、Batch_size 为 32、迁移优化策略选择 PaddlePaddle 封装的 AdamWeightDecayStrategy 策略、Weight_decay 设置为 0.01、Warmup 所占比重为 0.1、优化器选择 Adam、学习率均设置为 4e-5、分类器设置为全连接网络，使用激活函数 Softmax。实验数据第一列是文本内容，第二列为文本类别，列与列之间以 Tab 键分隔，以 tsv 格式输入，经过训练，不同联想值下各项指标如表 4 所示。观察看出当 n = 10 时，平均 F1 值为 0.807 3，达到最优，说明当为 SAO 扩展 10 个词汇后其语义表达能力达到最好。后续对比实验将基于 n = 10 展开。

表 4　BERT 模型不同联想值下各项分类指标

Id	BERT(n＝0)			BERT(n＝5)			BERT(n＝10)			BERT(n＝15)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
0	0.714 3	0.669 9	0.765 0	0.792 9	0.767 4	0.820 2	0.813 9	0.799 7	0.828 5	0.702 1	0.732 5	0.674 1
1	0.799 0	0.821 0	0.778 1	0.802 4	0.838 1	0.769 6	0.825 6	0.851 1	0.801 6	0.799 4	0.807 0	0.791 9
2	0.668 2	0.667 5	0.669 0	0.729 0	0.692 1	0.770 0	0.782 2	0.742 2	0.826 7	0.670 6	0.662 8	0.678 5
3	0.680 3	0.706 1	0.656 3	0.776 2	0.758 3	0.795 0	0.798 6	0.816 7	0.781 3	0.716 0	0.753 9	0.681 7
4	0.705 0	0.693 5	0.716 9	0.747 8	0.790 2	0.709 6	0.803 8	0.817 8	0.790 4	0.708 1	0.666 7	0.755 1
5	0.686 7	0.688 8	0.684 7	0.750 5	0.751 8	0.749 1	0.812 3	0.809 4	0.815 2	0.679 3	0.660 1	0.699 7
Average	0.708 9	0.707 8	0.711 7	0.766 5	0.766 3	0.768 9	0.806 1	0.806 1	0.807 3	0.712 6	0.713 8	0.713 5

4.4.2　不同学习率下的 SAO 短文本分类实验

学习率是影响分类指标的重要因素之一,不同学习率对训练过程的损失值产生不同影响:学习率过大容易造成梯度爆炸、损失步振幅难以平滑,导致模型无法收敛;学习率过小导致收敛速度缓慢,造成数据过拟合。本部分在联想值 n＝10 的基础上,将学习率(采用科学计数法)分别设置为 1e－6、2e－5、4e－4、4e－5 进行对比实验,其他配置参数不变。经过训练后,各项指标如表 5 所示:

表 5　BERT 模型不同学习率下各项分类指标

Id	1e－6			2e－5			4e－4			4e－5		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
0	0.597 1	0.532 2	0.679 9	0.819 4	0.818 8	0.820 0	0.719 9	0.711 8	0.728 1	0.853 8	0.826 4	0.883 1
1	0.691 4	0.673 8	0.709 9	0.847 2	0.867 5	0.827 9	0.753 6	0.788 1	0.721 9	0.860 4	0.874 2	0.847 0
2	0.481 7	0.452 2	0.515 2	0.817 1	0.816 4	0.817 7	0.667 2	0.641 4	0.695 2	0.828 3	0.816 5	0.840 4
3	0.317 0	0.429 0	0.251 4	0.800 0	0.828 7	0.773 2	0.655 7	0.683 0	0.630 4	0.821 9	0.840 7	0.803 9
4	0.468 6	0.439 8	0.501 4	0.783 4	0.745 9	0.824 8	0.669 2	0.608 0	0.744 1	0.832 3	0.826 2	0.838 6
5	0.442 4	0.587 9	0.354 7	0.781 0	0.781 0	0.781 0	0.666 7	0.751 8	0.598 9	0.829 3	0.849 8	0.809 8
Average	0.499 7	0.519 2	0.502 1	0.808 0	0.809 7	0.807 4	0.688 7	0.697 3	0.686 4	0.837 7	0.839 0	0.837 1

4.4.3　不同分类器下的 SAO 短文本分类实验

BERT 分类任务通常选用简单的全连接网络作为分类器,以 Softmax 作为激活函数实现自动分类。针对联想后的 SAO 文本这样多条语句表达相似含义的句式,选择其他网络模型作为分类器能否提高分类效果?为此,本部分在联想值 n＝10,学习率 learning-rate＝4e－5 的基础上对比全连接网络(Fully Connected Network,FC)和 LSTM 网络作为分类器对 P 值、R 值、F1 值的影响。全连接网络接受句子级别特征,输出对应[CLS]对应向量,格式为[－1,emb_size];设置为 LSTM 时,输出字符级别特征,结构为[－1,max_seq_len,emb_size],改变分类器时在 Task 添加一层网络即可。经过训练,各项指标见表 6。

表 6　BERT 模型不同分类器下各项分类指标

Id	FC			LSTM		
	P	R	F1	P	R	F1
0	0.870 9	0.869 4	0.872 5	0.779 5	0.780 1	0.779 0
1	0.875 6	0.883 6	0.867 7	0.811 2	0.804 6	0.817 9
2	0.846 8	0.863 0	0.831 2	0.733 3	0.739 5	0.727 1
3	0.835 5	0.827 4	0.843 9	0.752 3	0.765 0	0.740 0
4	0.847 4	0.822 1	0.874 3	0.731 0	0.723 1	0.739 0
5	0.838 1	0.853 5	0.823 3	0.757 4	0.756 7	0.758 1
Average	0.852 4	0.853 1	0.852 2	0.760 8	0.761 5	0.760 2

4.5　结果分析

4.5.1　不同分类模型实验结果对比分析

不同模型之间的平均 P 值、R 值、F1 值如图 6 所示,其中 BERT 取结果最高的一组作为对比。通过对比可以发现融合语义联想和 BERT 模型后的 SAO 平均 F1 值相较于 SVM 与 LSTM 更高,分别是 0.852 4、0.853 1、0.852 2;如图 7(a)(b)所示,BERT 在各个类别的 F1 值均处于最高水平,箱线图的类别分布也处于较为稳定的状态。因为 BERT 在大规模预训练语料的基础上,结合本文提出的语义联想方案,利用通用知识和领域知识相融合的方式,能够更显性地表征语义信息,同时语义降噪能够遗忘相关度较低、噪声较大的联想词,完备性进一步提高,从而分类效果更好。另外,SVM 相比 LSTM 识别效果较好,其中 F1 值达到 0.749 1,SVM 作为传统机器学习算法,能够有效地处理高维特征样本,同时在样本量较少的情况下的特征提取,相比较结

构复杂的深度学习模型，不需要完全依赖参数特征，单一领域的泛化能力较强。而深度学习模型 LSTM 对于数据的分类效果不如其他两个，因为 LSTM 作为循环神经网络的变形，模型参数和计算量更加复杂，需要输入大量数据学习不同类别之间的特征差异，对于数据不充足的实验容易丢失编码信息，SAO 短文本结构简洁，无需学习过长序列，词间数量也较为固定，构建的词典规模不大，因此无法发挥大规模参数计算的优势，导致分类效果不如其他模型。

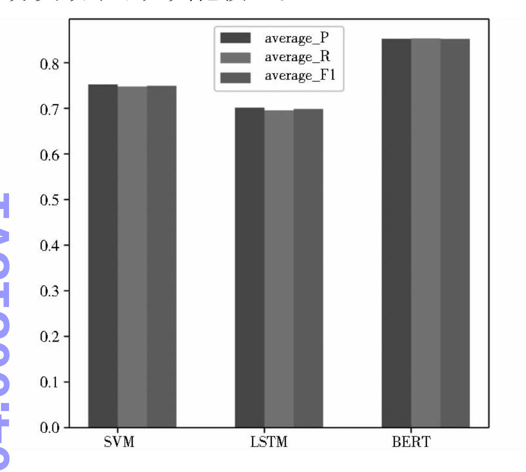


图 6　不同模型下各项平均指标比较

4.5.2　不同联想值的实验结果对比分析

不同联想值下的分类效果有所不同，如图 7（c）所示，分类结果 F1 值受 n 影响较大，当 n = 0，即不对 SAO 文本进行语义联想时，分类效果最差，维持在 0.65 – 0.75之间；当 n = 5 时，F1 值总体上升，说明扩展 5 个词后的效果比不扩展好；当 n = 10 时，各类别 F1 值提升效果显著，各类别维持在 0.8 上下，如图 7（d）所示，n = 10 时各类别差异最不显著，即差距最小，稳定性最好。但当 n 值达到 15 时，F1 值大幅下跌，降低到 0.67 – 0.75 上下，F1 最大值和平均值之间的差异比较大，分类效果差距比较大，稳定性较差，这是因为 Word2Vec 训练数据规模有限，无法完全为每个词语匹配出最相似的表达，因此当 n = 15 时，联想词的整体关联性下降，导致 SAO 文本与对应类别的偏移度上升，分类效果降低。从 0 到 5 到 10，随着联想值得提高，分类 F1 值随之提高，说明随着扩展词的增加，SAO 短文本语义信息越丰富，每一个类别之间的差异更加显著，达到联想值达到 15 时，效果下降，可以认为 SAO 短文本分类效果随着联想值提高而提高，但联想值需要控制在局部范围内。

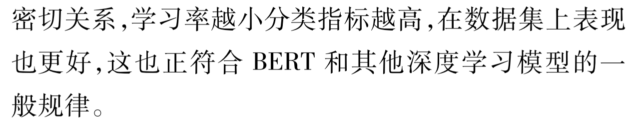
(a) 不同模型下各类别F1值　(c) 不同扩展值下各类别F1值　(e) 不同学习率下各类别F1值　(g) 不同分类器下各类别F1值

(b) 不同模型下各类别F1值分布　(d) 不同扩展值下各类别F1值分布　(f) 不同学习率下各类别F1值分布　(h) 不同分类器下各类别F1值分布

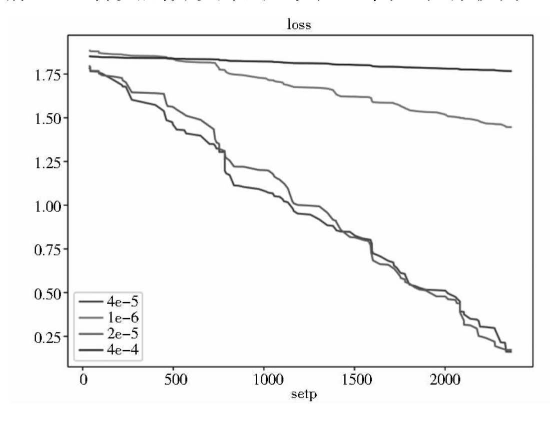
图 7　不同变量对分类结果的影响

4.5.3　不同学习率和分类器的实验结果对比分析

不同学习率下的损失与精确度变化如图 8 所示，当学习率设置为 1e – 6 和 4e – 4 时，训练过程的损失下降缓慢，精确度最高在 0.6 左右，随着学习率减小，训练损失逐渐下降、精度值逐渐提升，当学习率为 4e – 5 时，达到四项最优化，各类别 F1 值及其分布如图 7



(e)、(f)所示,F1 值在 2e－5 和 4e－5 时最为集中、趋于稳定,其余则较为分散、F1 值离散程度较高。可以看出,图情 SAO 各类别分类效果与学习率大小有较为密切关系,学习率越小分类指标越高,在数据集上表现也更好,这也正符合 BERT 和其他深度学习模型的一般规律。



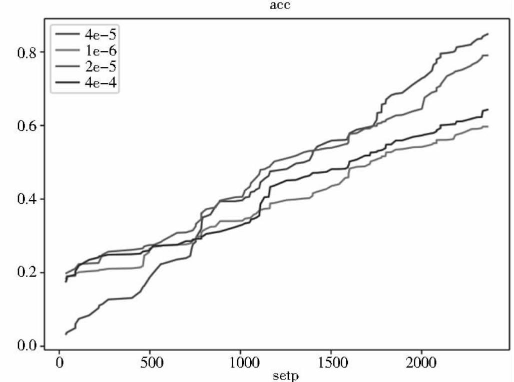


图 8　不同学习率对结果的影响

图 7(g)、(h)中不同分类器下各类别指标采取在 BERT 后连接全连接层和 LSTM 层的差异显著,只使用全连接层的平均 F1 值达到 0.852 2,各类别 F1 值整体维持在 0.85 左右,但接入 LSTM 后再使用激活函数 Softmax 分类的结果只有 0.760 2,与基线 LSTM 相同,BERT 后连接 LSTM 再进行分类的向量计算并没有提高参数增加带来的优势,效果反而不如直接连接的全连接层。

综上所述,对比不同模型、不同联想值、学习率、分类器对 SAO 分类的效果后,表明与传统机器学习、深度学习相比,融合语义联想和 BERT 的 SAO 短文本分类方法有更显著的优势,相比较单纯的 SVM、LSTM 和 BERT 分类模型,F1 值分别提高了 0.103 1、0.153 8 和 0.140 5,在局部范围内分类效果与联想词数量呈正相关关系,在联想值固定情况下,学习率和分类器对结果也有一定影响,最终当联想值为 10,学习率为 4e－5 时 SAO 分类效果达到最优化。

5　结语

针对 SAO 短文本分类存在的问题,本文提出了融合语义联想和 BERT 的 SAO 短文本分类方法,以期延伸 SAO 表征范围、提高了融合学习率,并采用该分类方法对图情领域 SAO 进行了实证研究,通过对实验结果的对比分析,发现融入图情专业知识的输入数据结合 BERT 能够更好地识别图情领域 SAO 短文本,证明“语义联想 + BERT”的 SAO 短文本分类方法是可行的。但是本文所提出的方法还存在一定的局限性,由于语义联想模型语料受限的原因,所以实验无法为每个 SAO 短文本进行充分地语义联想,由此出现了当联想值进一步提高时分类效果降低的现象,此外,该方法也未能扩展到更多领域进行适应性检测,在接下来的研究工作中,笔者将对上述问题做进一步探究。

参考文献:

[1] CASCINI G, FANTECHI A, SPINICCI E. Natural language processing of patents and technical documentation[C]//International workshop on document analysis systems. Berlin: Springer, 2004: 508－520.

[2] CHOI S, PARK H, KANG D, et al. An sao-based text mining approach to building a technology tree for technology planning[J]. Expert systems with applications, 2012, 39(13): 11443－11455.

[3] WANG X, WANG Z, HUANG Y, et al. Identifying r&d partners through subject-action-object semantic analysis in a problem & solution pattern[J]. Technology analysis & strategic management, 2017, 29(10): 1167－1180.

[4] TSOURIKOV V M, BATCHILO L S, SOVPEL I V. Document semantic analysis/selection with knowledge creativity capability utilizing subject-action-object (sao) structures: U. S. Patent 6,167,370[P]. 2000－12－26.

[5] 付芸,汪雪锋,李佳,等. 基于 SAO 结构的创新解决方案遴选研究——以空气净化技术为例[J]. 图书情报工作,2019,63(6):75－84.

[6] 许海云,王振蒙,胡正银,等. 利用专利文本分析识别技术主题的关键技术研究综述[J]. 情报理论与实践,2016,39(11):131－137.

[7] 胡正银,刘春江,隗玲,等,文奕. 面向 TRIZ 的领域专利技术挖掘系统设计与实践[J]. 图书情报工作,2017,61(1):117－124.

[8] 杨超,朱东华,汪雪锋,等. 专利技术主题分析:基于 SAO 结构的 LDA 主题模型方法[J]. 图书情报工作,2017,61(3):86－96.

[9] CHANG P L, WU C C, Leu H J. Using patent analyses to monitor

the technological trends in an emerging field of technology: a case of carbon nanotube field emission display[J]. Scientometrics, 2010, 82(1): 5 – 19.

[10] GUO J, WANG X, LI Q, et al. Subject-action-object-based morphology analysis for determining the direction of technological change[J]. Technological forecasting and social change, 2016, 105: 27 – 40.

[11] LI X, WANG J J, YANG Z. Identifying emerging technologies based on subject-action-object[J]. Journal of intelligence, 2016, 35(3): 80 – 84.

[12] 王晓宇,苗红,王芳.技术知识的跨领域应用及潜在技术方案的识别[J].图书情报工作,2016,60(23):87 – 96.

[13] 胡正银,方曙,张娴,等.个性化语义TRIZ构建研究[J].图书情报工作,2015,59(7):123 – 131.

[14] MANEK A S, SHENOY P D, MOHAN M C, et al. Aspect term extraction for sentiment analysis in large movie reviews using gini index feature selection method and SVM classifier[J]. World Wide Web, 2017, 20(2): 135 – 154.

[15] BACHHETY S, DHINGRA S, JAIN R, et al. Improved multinomial naïve bayes approach for sentiment analysis on social media [J]. International journal of information systems & management science, 2018, 1(1).

[16] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257 – 286.

[17] BREIMAN L. Random forests[J]. Machine learning, 2001, 45 (1): 5 – 32.

[18] 高金勇,徐朝军,冯奕競.基于迭代的TFIDF在短文本分类中的应用[J].情报理论与实践,2011,34(6):120 – 122.

[19] 范云杰,刘怀亮.基于维基百科的中文短文本分类研究[J].现代图书情报技术,2012(3):47 – 52.

[20] MINAEE S, KALCHBRENNER N, CAMBRIA E, et al. Deep learning based text classification: a comprehensive review[J]. arXiv preprint arXiv:2004.03705, 2020.

[21] YIN W, KANN K, YU M, et al. Comparative study of cnn and rnn for natural language processing[J]. arXiv preprint arXiv: 1702.01923, 2017.

[22] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735 – 1780.

[23] OLAH C. Understanding lstm networks[EB/OL][2015 – 8 – 27]. https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[24] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[25] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998 – 6008.

[26] 邓三鸿,傅余洋子,王昊.基于LSTM模型的中文图书多标签分类研究[J].数据分析与知识发现,2017,1(7):52 – 60.

[27] 吕璐成,韩涛,周健,等.基于深度学习的中文专利自动分类方法研究[J].图书情报工作,2020,64(10):75 – 85.

[28] LEE J, DERNONCOURT F. Sequential short-text classification with recurrent and convolutional neural networks[J]. arXiv preprint arXiv:1603.03827, 2016.

[29] 秦成磊,章成志.基于层次注意力网络模型的学术文本结构功能识别[J].数据分析与知识发现,2020,4(11):26 – 42.

[30] 陶志勇,李小兵,刘影,等.基于双向长短时记忆网络的改进注意力短文本分类方法[J].数据分析与知识发现,2019,3(12):21 – 29.

[31] 余本功,朱梦迪.基于层级注意力多通道卷积双向GRU的问题分类研究[J].数据分析与知识发现,2020,4(8):50 – 62.

[32] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[33] SUN C, QIU X, XU Y, et al. How to fine-tune bert for text classification? [C]//China national conference on Chinese computational linguistics. Cham: Springer, 2019: 194 – 206.

[34] LEE J S, HSIANG J. Patentbert: Patent classification with fine-tuning a pre-trained bert model[J]. arXiv preprint arXiv:1906.02124, 2019.

[35] LU X, NI B. BERT – CNN: A hierarchical patent classifier based on a pre-trained language model[J]. arXiv preprint arXiv:1911.06241, 2019.

[36] 刘欢,张智雄,王宇飞.BERT模型的主要优化改进方法研究综述[J/OL].数据分析与知识发现:1 – 17[2021 – 01 – 05]. https://doi.org/10.11925/infotech.2096-3467.2020.0965.

[37] LIU W, ZHOU P, ZHAO Z, et al. K-BERT: Enabling Language Representation with Knowledge Graph[J]. arXiv preprint arXiv:1909.07606, 2019.

[38] YU S, SU J, LUO D. Improving BERT-based text classification with auxiliary sentence and domain knowledge[J]. IEEE access, 2019, 7: 176600 – 176612.

[39] ORKPHOL K, YANG W. Word sense disambiguation using cosine similarity collaborates with Word2vecand WordNet[J]. Future Internet, 2019, 11(5): 114.

[40] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[41] 中国大百科全书总编辑委员会.中国大百科全书 图书馆学·情报学·档案学[M].北京:中国大百科全书出版社, 2002.

[42] 丘东江. 新编图书馆学情报学辞典[M].北京:科学技术文献出版社, 2006.

[43] 白如江,张庆芝,孙一钢.科技文献知识基因表达及遗传与变异研究[J].图书情报工作,2020,64(4):78 – 87.

[44] 图书馆·情报与文献学名词审定委员会. 图书馆·情报与文献学名词2019[M].北京:科学出版社,2019.

[45] ASHKAN J, HAMED E, MIHAN H, et al. Improvement in auto-

matic classification of Persian documents by means of support vector machine and representative vector[C]//International conference on innovative computing technology. Berlin: Springer, 2011: 282－292.

[46] 杨敏,谷俊. 基于 SVM 的中文书目自动分类及应用研究[J]. 图书情报工作,2012,56(9):114－119.

[47] 王东波,何琳,黄水清. 基于支持向量机的先秦诸子典籍自动分类研究[J]. 图书情报工作,2017,61(12):71－76.

[48] WANG J H, LIU T W, LUO X, et al. An LSTM approach to short text sentiment classification with word embeddings[C]//Proceedings of the 30th conference on computational linguistics and speech processing (ROCLING 2018). Hsinchu: ACLCLP, 2018: 214－223.

作者贡献说明：

张玉洁：实验实施,论文撰写；

白如江：研究选题与设计,论文修改和审阅；

刘明月：实验方案设计,论文修改；

于纯良：数据语料处理,论文修改。

Research on SAO Short Text Classification in LIS Based on Semantic Association and BERT

Zhang Yujie[1] Bai Rujiang[1] Liu Mingyue[1] Yu Chunliang[2]

[1] Institute of Information Management, Shandong University of Technology, Zibo 255049

[2] Yantai University Library, Yantai 264005

Abstract: [**Purpose/significance**] Aiming at the shortage of semantic features and insufficient domain knowledge in the classification of SAO structure short texts, this paper proposes a SAO classification method combining semantic association and BERT in order to improve the classification effect. [**Method/process**] Taking the SAO short text in the library and information science field as the data source, firstly, a semantic association scheme including the three links of "Expansion-Reconstruction-NoiseReduction" was designed. The semantic information of SAO was extended through semantic expansion and SAO reconstruction, and the extended noise interference problem was solved by semantic noise reduction; then used the BERT model to train the SAO short text after semantic association; finally realized automatic classification in the classification part. [**Result/conclusion**] After comparing different association values, learning rates and classifiers, the experimental results show that when the association value is 10 and the learning rate is 4e－5, the SAO short text classification effect is optimal, and the average F1 value is 0.852 2, which is comparable to SVM and LSTM compared with pure BERT, the F1 value is increased by 0.103 1, 0.153 8 and 0.140 5 respectively.

Keywords: SAO short text classification semantic association BERT